Homework 2

ST 697 / Fall 2017

Due: Tues Sept 26

Problem 2.1: Centering and Scaling

Suppose we use training data $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^n$ where $y_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^p$ to estimate the parameters $\beta = (\beta_0, \ldots, \beta_p)$ according to least squares (assume $p + 1 < n$), i.e.,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

Also consider alternative predictors that are centered and scaled versions of the original predictors:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$ and $s_j^2 = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$.

a. What is $\sum_{i=1}^n z_{ij}$ and $\sum_{i=1}^n z_{ij}^2$?

b. Show how would you can create the new predictor variables in the software of your choice. Check that the values match what is obtained in part (a). Show code.

c. Derive the optimal parameters (as functions of the originals) if the alternative predictors are used. I.e., find

$$\hat{b} = \underset{b \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} z_{ij} b_j \right)^2$$

d. Given that you have $\hat{b}$, derive the optimal $\hat{\beta}$ as functions of $\hat{b}$.

e. If we use the alternative predictors and also replace y with the centered version $y_i' = y_i - \bar{y}$, do we save one degree of freedom? Explain why or why not?

Problem 2.2 Ensemble of linear models

A linear *ensemble* model, $m_w(\cdot)$, is of the form:

$$m_w(x) = \sum_{j=1}^{J} w_j m_j(x)$$

where $w \in \mathbb{R}^J$ is a vector of weights and $m_j(x)$ is the predicted output from model j at x.

a. If there are p total predictors and all the models are *linear models* (e.g., $m_j(x, \beta) = \beta_{j,0} + \sum_{k=1}^p x_k \beta_{j,k}$), show that $m_w(x)$ is also a linear model and provide the coefficient values. Note: if model j does not use predictor x_k, then $\beta_{j,k} = 0$.

b. Suppose we have used the training data to fit J models $\{m_j(\cdot)\}_{j=1}^J$. What is wrong with using the training data to find the optimal weights in a least squares sense, i.e.

$$w^* = \underset{w}{\arg\min} \sum_{i=1}^{n} \left(y_i - m_w(x_i) \right)^2$$

Describe the optimal weights? Will they be of special form? Give a suggested improvement.

Problem 2.3 AIC for linear regression

Suppose we have n iid observations $D = \{(x_i, y_i)\}_{i=1}^n$ from the model $Y_i = X_i^\mathsf{T}\beta + \epsilon_i$, where $\epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

a. Write out the log-likelihood: $\log L(\beta, \sigma \mid D)$ as functions of β and σ.

b. Find the MLE
$$(\hat{\beta}, \hat{\sigma}^2) = \arg\max_{\beta,\sigma} \log L(\beta, \sigma \mid D)$$

c. Derive the equation for AIC. Reduce as much as possible.

d. Suppose we have two *nested* linear models, model 1 has p parameters and model 2 has the same p parameters as model 1 plus d additional parameters. State the necessary conditions for model 2 being favored with respect to i) AIC ii) R^2 iii) adjusted R^2.

Problem 2.4 (ESL 3.12) Data Augmentation for Ridge Regression

Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set. We augment the centered matrix X with p additional rows $\sqrt{\lambda} I_p$, and augment y with p zeros.

Problem 2.5 Ridge Regression Algorithm

a. Write a function for ridge regression. Inputs should be X, y, and a sequence of λ values. Outputs should be the estimated coefficients and edof (for each λ), and anything else necessary for part b.
   - Should the intercept be penalized?
   - Should the input variables be standardized?

b. Write another function that will make predictions given X' and a single λ value.

Problem 2.6 Prediction Contest: Real Estate Pricing

This problem uses the realestate-train and realestate-test (click on links for data).

The goal of this contest is to predict sale price (in thousands) (`price` column) using ridge regression. Evaluation of the test data will be based on the root mean squared error RMSE $= \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}$ for the n test set observations.

a. Fit a collection of ridge regression models (indexed by penalty value λ) to the training data (`realestate-train.csv`). You can use an existing function (e.g., `glmnet::glmnet()` or `MASS::lm.ridge()`) or your own from 2.5. Note: there are some categorical predictors.

b. Plot the estimated coefficient values (i.e., ridge trace) against: (i) root mean squared error, (ii) λ (or $\log\lambda$ if it looks better), (iii) penalty $P(\beta) = \sum_j |\beta_j|^2$ (not including intercept), and (iv) the effective degrees of freedom, $df(\lambda)$. Comment on the results.

c. Estimate the optimal shrinkage parameter, λ^*, using GCV and LOOCV. Report the shrinkage parameter that each method selects.

d. Make a decision about the best value of λ and predict the response for the testing data (`realestate-test.csv`). Notice the test data does not have the `price` column, this is what you are predicting. Email me a one column `.csv` file with your predictions.
   - Competitive advantage: you are free to use any technique to estimate the optimal λ.

e. Report the anticipated performance of your method in terms of RMSE. We will see how close your performance assessment matches the actual value.